conformations. A term that adjusts the model to this bimodal distribution could be expressed as follows, with all distances in lattice units.

$$E_{struct} = \Sigma\, Es(i) \tag{4a}$$

with:

$Es(i) = -2\varepsilon_{gen}$,

$$\text{for } r^2_{i,i+4} < 33 \text{ and } (\mathbf{v}_i \ \ \mathbf{v}_{i+3}) > 0 \tag{4b}$$

or

$Es(i) = -2\varepsilon_{gen}$,

$$\text{for } 48 < r^2_{i,i+4} < 145 \text{ and } (\mathbf{v}_{i+1} \ \ \mathbf{v}_{i+2}) < 0. \tag{4c}$$

The first set of conditions (equation 4(b)) describes a loosely defined, helical conformation, while the second (equation 4(c)) describes an extended, β-type fragment. Thus, equation 4(b) states that the distance between the i-th and i + 4th side chain in a helix has to be small (here, below about 8 Å). The second condition states that the chain has to make a slight turn. A corresponding set of conditions is defined for β-type expanded states. In both cases, the cut-off distances and the angular restrictions are selected in a very permissive way based on the observed distributions for native proteins. The permissive definition of local conformational biases drives the model system towards a loosely defined protein-like chain geometry, yet it still allows substantial local mobility. As mentioned before, in preferred simulations, the value of ε_{gen} has been assumed to be equal to 1 k_BT.

"Hydrogen bonds" and generic packing biases

Model hydrogen bonds provide similar structure-regularizing biases with respect to tertiary interactions, as do the generic short-range interactions for secondary structural regularities. Residue i is considered to be hydrogen-bonded to residue j when the orthogonal vector $\mathbf{w}_i$ (originating from the bead i) touches any of

the 17 points of the excluded volume cluster of residue j. In various embodiments of the model, two hydrogen bonds originate from a given residue. The geometry of hydrogen bonds is depicted in Figure 5. Only residues that are "in contact" could be hydrogen-bonded. That is, there is the same long-range cut-off for side group pair interactions as for hydrogen bonding. The energy of the hydrogen bond network is defined as follows:

$$E_{H\text{-}bond} = -\varepsilon_{H\text{-}bond} \Sigma (\delta^{+} + \delta^{-} + \delta^{+,-}) \quad (5)$$

where δ^{+}, δ^{-}, $\delta^{+,-}$ are equal to 1 when the "right handed," the "left handed," and both hydrogen bonds originating from residue i are satisfied, respectively. Otherwise, the corresponding terms are equal to zero. The last term, $\delta^{+,-}$, is a cooperative hydrogen bond energy gained only upon local saturation. The numerical value of this parameter was assumed to be equal to about 1.0-1.25 k_BT. Values of this parameter toward the lower end of the range tend to accelerate folding, while values toward the higher end tend to build structures of slightly better quality. In any event, these effects are small, and it is preferred to use a term having the same value (1.0) in all isothermal Monte Carlo runs used for energy comparisons.

Two other generic terms that enforce protein-like packing regularities also have been introduced. The first one is a "contact map propagator" that reflects the most common patterns seen in all side chain contact maps of globular proteins.[18] It is defined in the following way:

$$E_{map} = -\varepsilon_{gen}(\Sigma\Sigma(\delta_{i,j} \cdot \delta_{i+1,j+1} \cdot \delta_{i-1,j-1})\delta_{par} + \Sigma\Sigma(\delta_{i,j} \cdot \delta_{i-1,j+1} \cdot \delta_{i+1,j-1})\, \delta_{apar}) \quad (6)$$

where δ_{ij} id equal to 1 (0) when residues i and j are (not) in contact. δ_{par} is equal to 1 only when the corresponding chain fragments are oriented in a parallel fashion, *i.e.*, $(v_{i-1} + v_i)$ x $(v_{j-1} + v_j)$. Similarly , δ_{apar} is equal to 1 when the chain fragments are anti-parallel. In the above equation and in equation 7, below, $\varepsilon_{gen} = 1$ is the same parameter as the one used in the short-range generic terms.

A second packing regularizing term provides an additional cohesive energy between secondary structure elements by favoring the parallel packing of pairs of hydrophilic residues and the anti-parallel packing of pairs of hydrophobic residues. Consequently, since it exploits sequence information, this term is not purely generic; however, it is reduced to a two-letter (HP) code.

$$E_{packing} = -\varepsilon_{gen} \Sigma\Sigma(\delta_{PP} \cdot \delta_{pp} + \delta_{HH} \cdot \delta_{app}) \quad (7)$$

where δ_{PP} (δ_{HH}) is equal to 1 when both residues in contact are hydrophilic, P, (hydrophobic, H), according to the Kyte-Doolittle hydrophobicity scale.[19] The value of δ_{pp} is equal to 1 only when the packing of the side chain pair is parallel; *i.e.*, $(v_{i-1} - v_i) \times (v_{j-1} - v_j)0$. Similarly, δ_{app} is equal to 1 only when the packing of the side chain pair is parallel; *i.e.*, $(v_{i-1} - v_i) \times (v_{j-1} - v_j)0$.

Various structure regularizing terms described in this and the previous section reflect the various structural regularities seen in globular proteins. Each term accounts for a different correlation that could be easily detected by statistical analysis of the geometry of the side-chain-only representation of protein structures. Except for the last term (which depends on some sequence features), they are sequence independent: the underlying regularities are true for all types of structural motifs of globular proteins. During Monte Carlo simulations, these generic potentials provide a very strong bias against nonsensical, non-protein like conformations. Such conformations would otherwise be quite frequent due to the reduced character of the protein representation. In the presence of these generic contributions to the model force field, the requirements for sequence-specific potentials are lower; they have to select between various protein-like confirmations, which makes the selection easier (and computationally less expensive) than in the much broader conformational space of an unrestricted model chain.

Sequence-specific long-range interactions

These interactions are defined as follows: